Step Forward Cross Validation for Bioactivity Prediction: Out of Distribution Validation in Drug Discovery

Udit Surya Saha, Michele Vendruscolo,
*Anne E. Carpenter, Shantanu Singh, Andreas Bender, Srijit Seal**

seal@broadinstitute.org

Step-Forward Cross-Validation

Recent advances in machine learning for materials science have inspired the adaptation of validation methods for drug discovery.

Traditional random split cross-validation often fails to generalize well for out-of-distribution data.

We propose a k-fold n-step forward cross-validation (SFCV) approach to improve model performance on predicting small molecule bioactivity, enhancing the real-world applicability of these models.

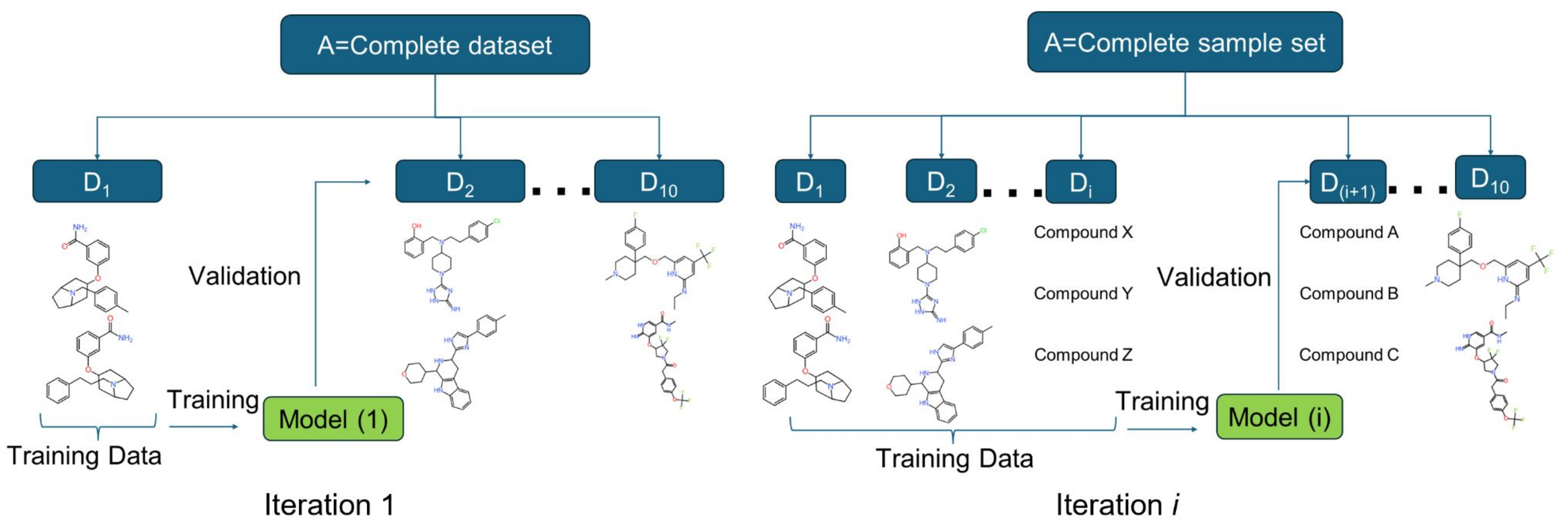

Figure 1: Workflow for 10-fold step-forward cross-validation (SFCV). The Di dataset block can be sorted via a calculated or experimental molecular property. Here, we used logP.

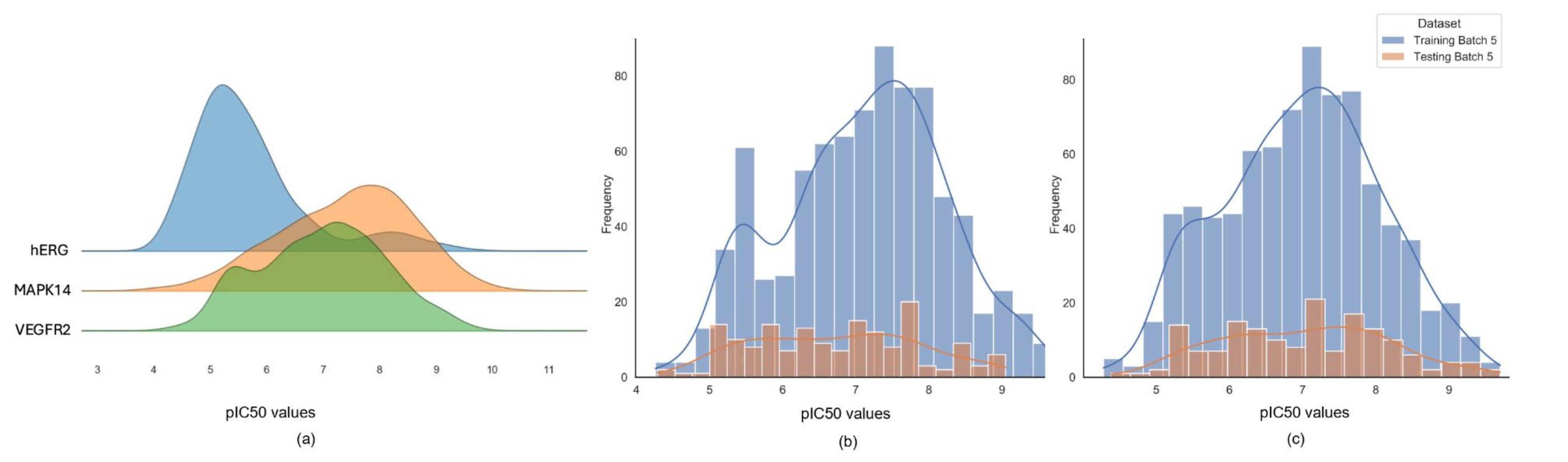

Figure 2: (a) Distribution of biological activity (pIC50 values) for the three protein target datasets. (b,c) Distribution of pIC50 values in the training set and test set of the 5th iteration for sorted SFCV (b) and unsorted SFCV (c). SFCV: Step-Forward Cross-Validation

Validation Methods Comparison

Four validation methods were compared: sorted SFCV, unsorted SFCV, cross-validation with random splits, and cross-validation with scaffold splits.

Sorted SFCV selected test compounds with progressively lower logP values, aligning with real-world drug optimization processes.

Scaffold-based splits and random splits did not show consistent trends in selecting test compounds, often leading to suboptimal evaluation of model performance.

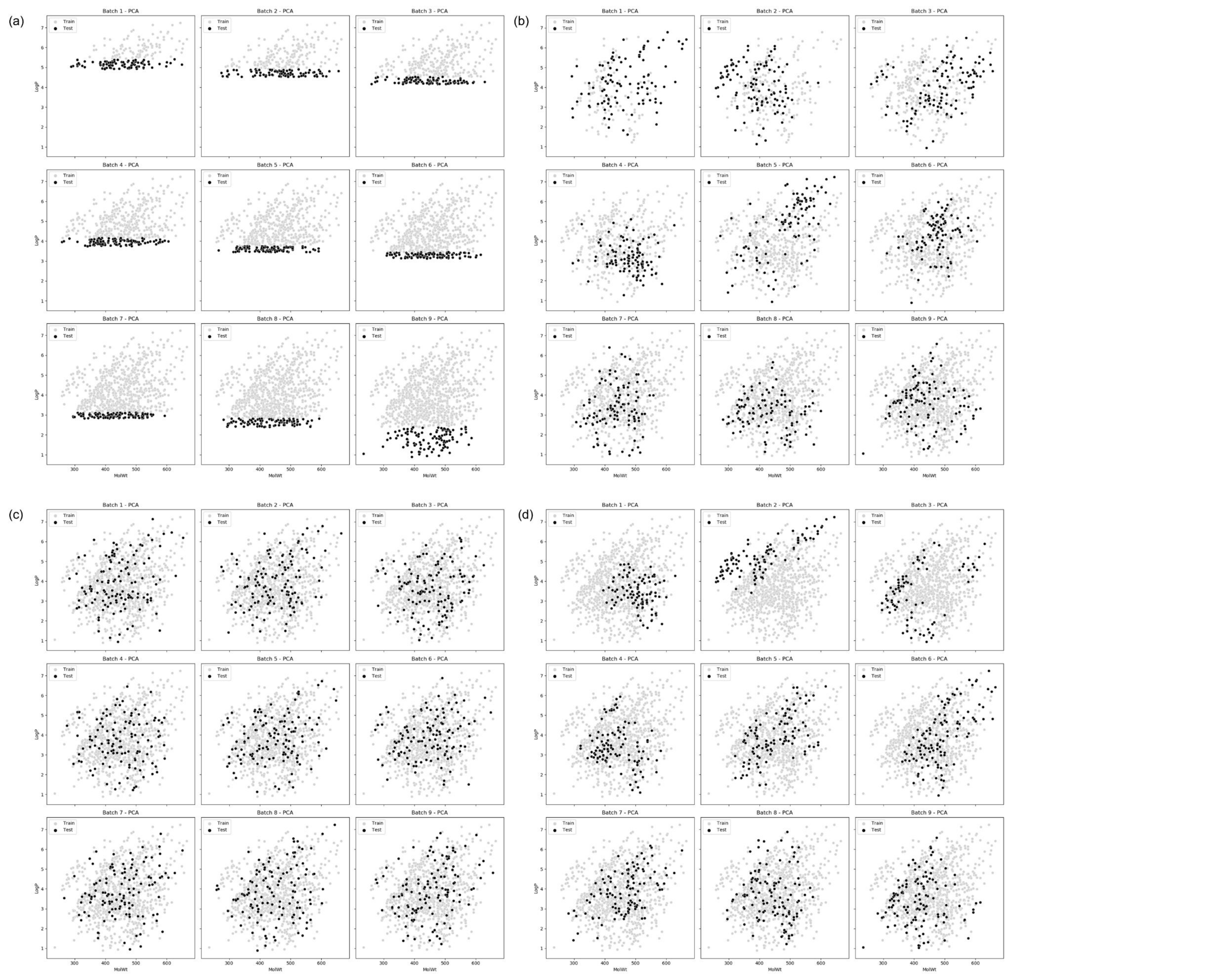

Figure 3: Comparison of logP and Molecular Weight for physico-chemical space for compounds selected as training and test sets across various iterations for the hERG target prediction task for (a) sorted SFCV, (b) unsorted SFCV, (c) cross-validation with random splits, and (d) cross-validation with scaffold splits for the (first) nine iterations for Random Forest models. SFCV: Step-Forward Cross-Validation

Performance Metrics

Sorted SFCV demonstrated a higher ability to identify structurally novel compounds compared to other methods.

It also showed more accurate predictions for discovery compounds—compounds with desirable bioactivity (pIC50 < 5.2 for hERG). Although sorted SFCV had a higher absolute error for some compounds, it provided a more challenging and realistic evaluation of model performance in drug discovery.

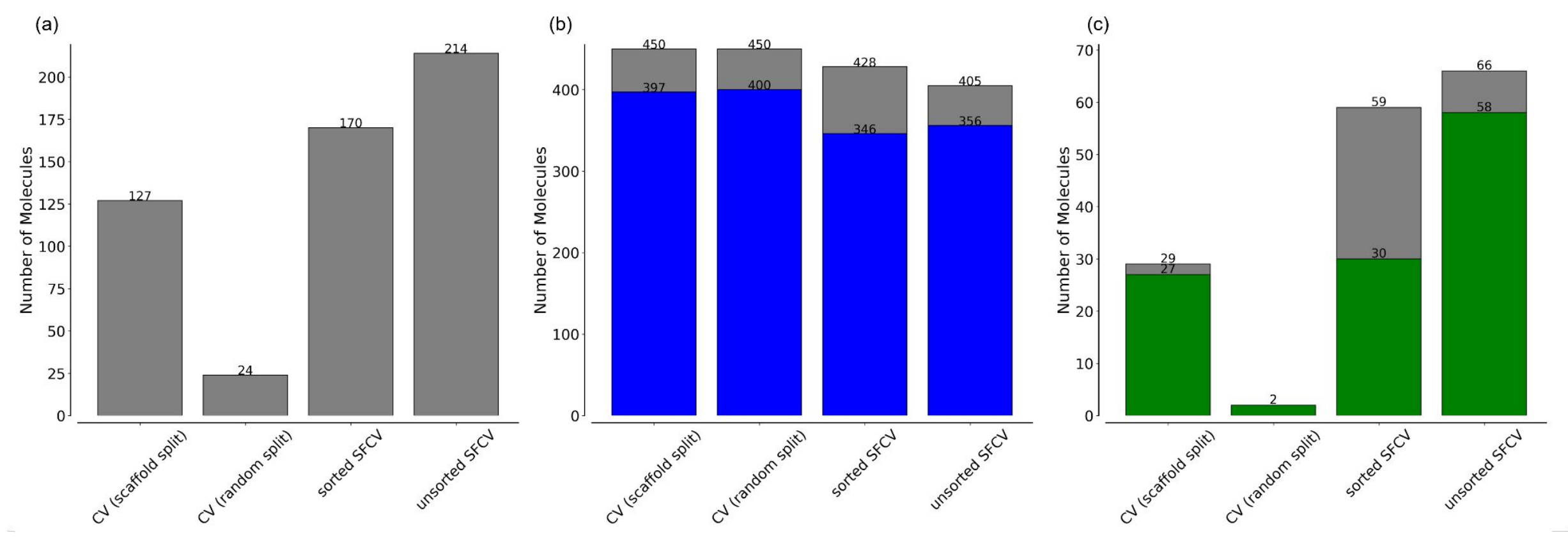

Figure 4: (a) The total number of compounds dissimilar to training data (Tc<0.55). (b) The total number of discovery compounds (pIC50<5.2) in the test set and the colored stacks show how many predictions are within a 0.5 log unit error range. (c) The number of discovery compounds dissimilar to training data (Tc<0.55) and the colored stacks show how many predictions are within a 0.5 log fold unit error range. Results refer to the hERG target prediction task across four validation methods: sorted SFCV, unsorted SFCV, cross-validation with random splits, and cross-validation with scaffold splits (combined for all test folds). Tc: Tanimoto Similarity; SFCV: Step-Forward Cross-Validation.

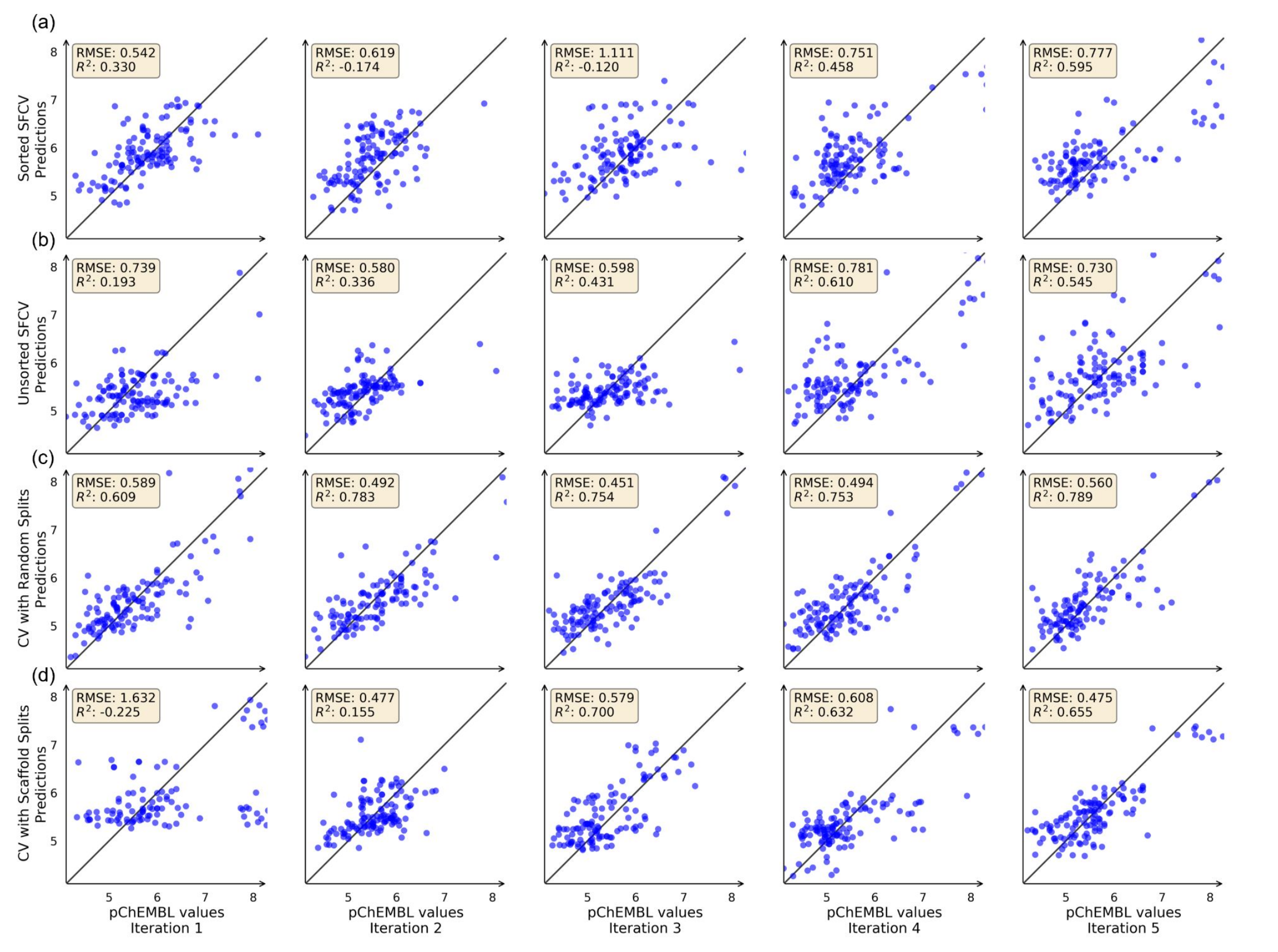

Figure 5: Parity plots for hERG target prediction for (a) sorted SFCV, (b) unsorted SFCV, (c) cross-validation with random splits, and (d) cross-validation with scaffold splits for the first five iterations for Random Forest models. SFCV: Step-Forward Cross-Validation

Discovery Yield and Novelty Error

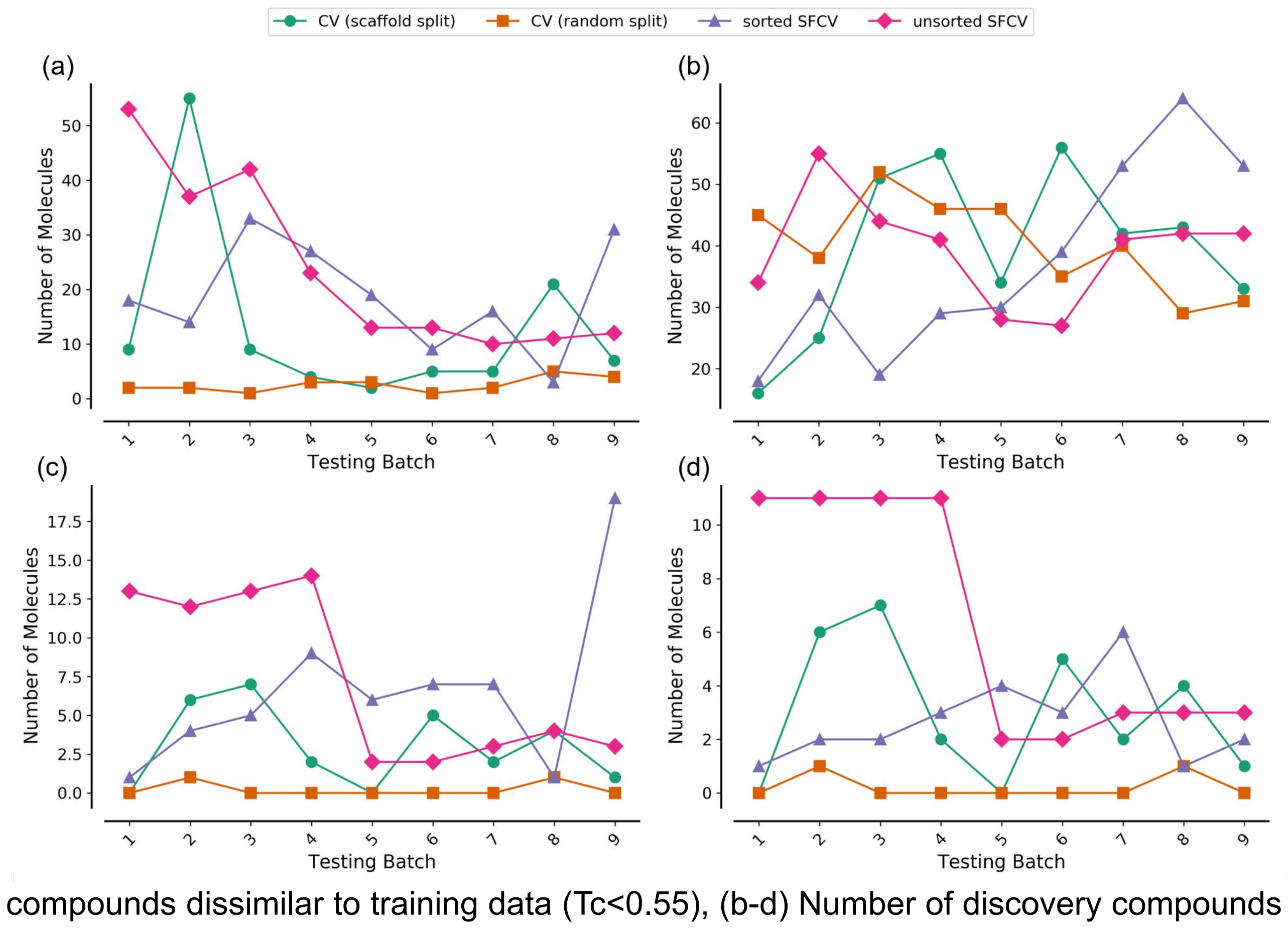

Figure 6: (a) Number of compounds dissimilar to training data (Tc<0.55), (b-d) Number of discovery compounds (pIC50<5.2) in the test set predicted within an error range of 0.5 log unit (b), and discovery compounds dissimilar to training data (Tc<0.55) (c), and discovery compounds dissimilar to training data (Tc<0.55) correcting predicted within an error range of 0.5 log unit (d), as shown for the hERG target prediction task across four validation methods of sorted SFCV, unsorted SFCV, cross-validation with random splits, and cross-validation with scaffold splits (for each of the first nine test folds). Tc: Tanimoto Similarity; SFCV: Step-Forward Cross-Validation.

Discovery yield, defined as the fraction of discovery compounds accurately predicted within an error range of 0.5 log units, was more consistent in sorted SFCV. Novelty error, the mean absolute error for structurally novel compounds, remained low and consistent in sorted SFCV. This indicates that sorted SFCV minimizes overfitting and improves generalization to novel chemical spaces.

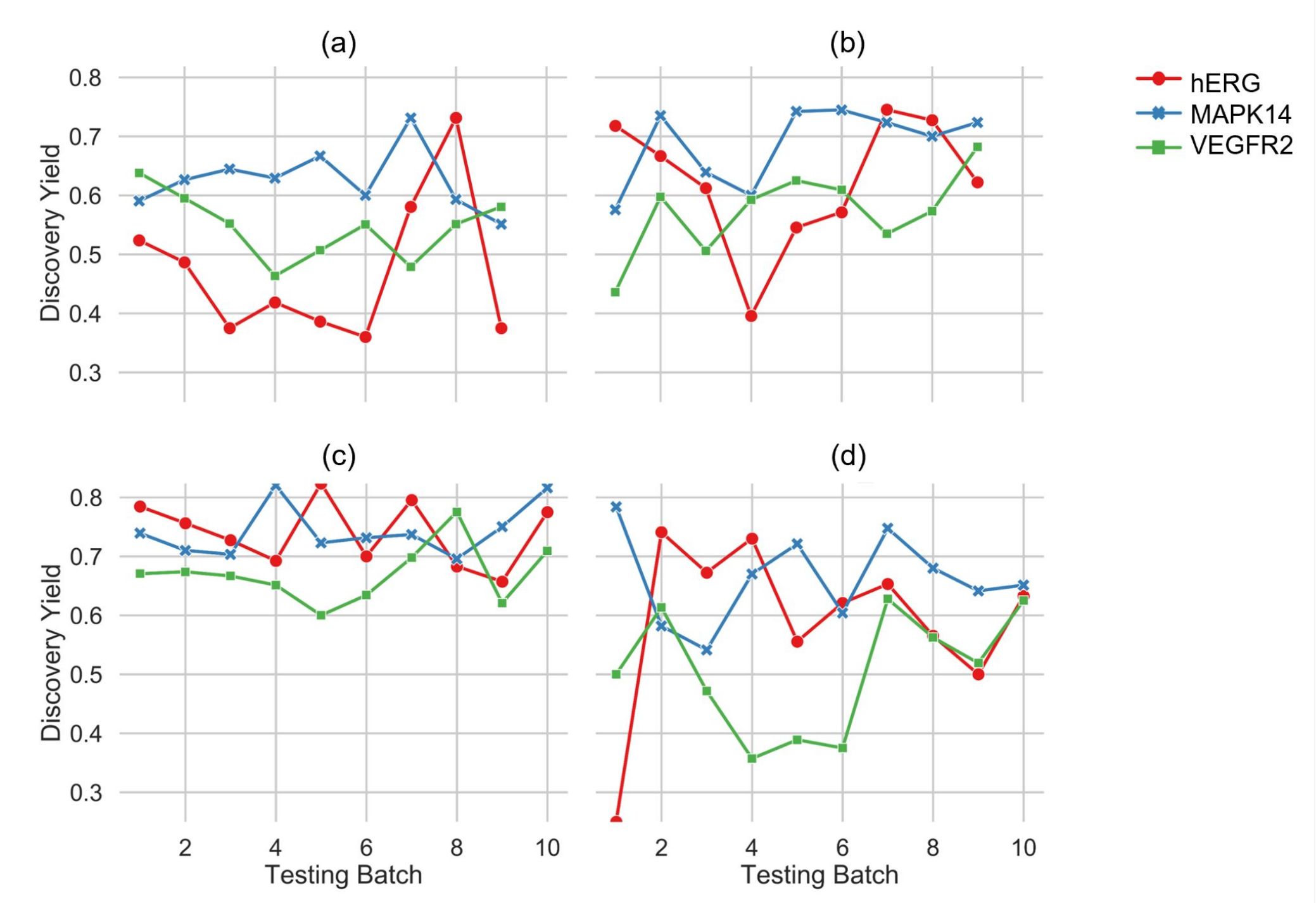

Figure 7: Discovery yield for models validated using (a) sorted SFCV, (b) unsorted SFCV, (c) cross-validation with random splits, and (d) cross-validation with scaffold splits for all three protein targets when using a Random Forest model. SFCV: Step-Forward Cross-Validation.

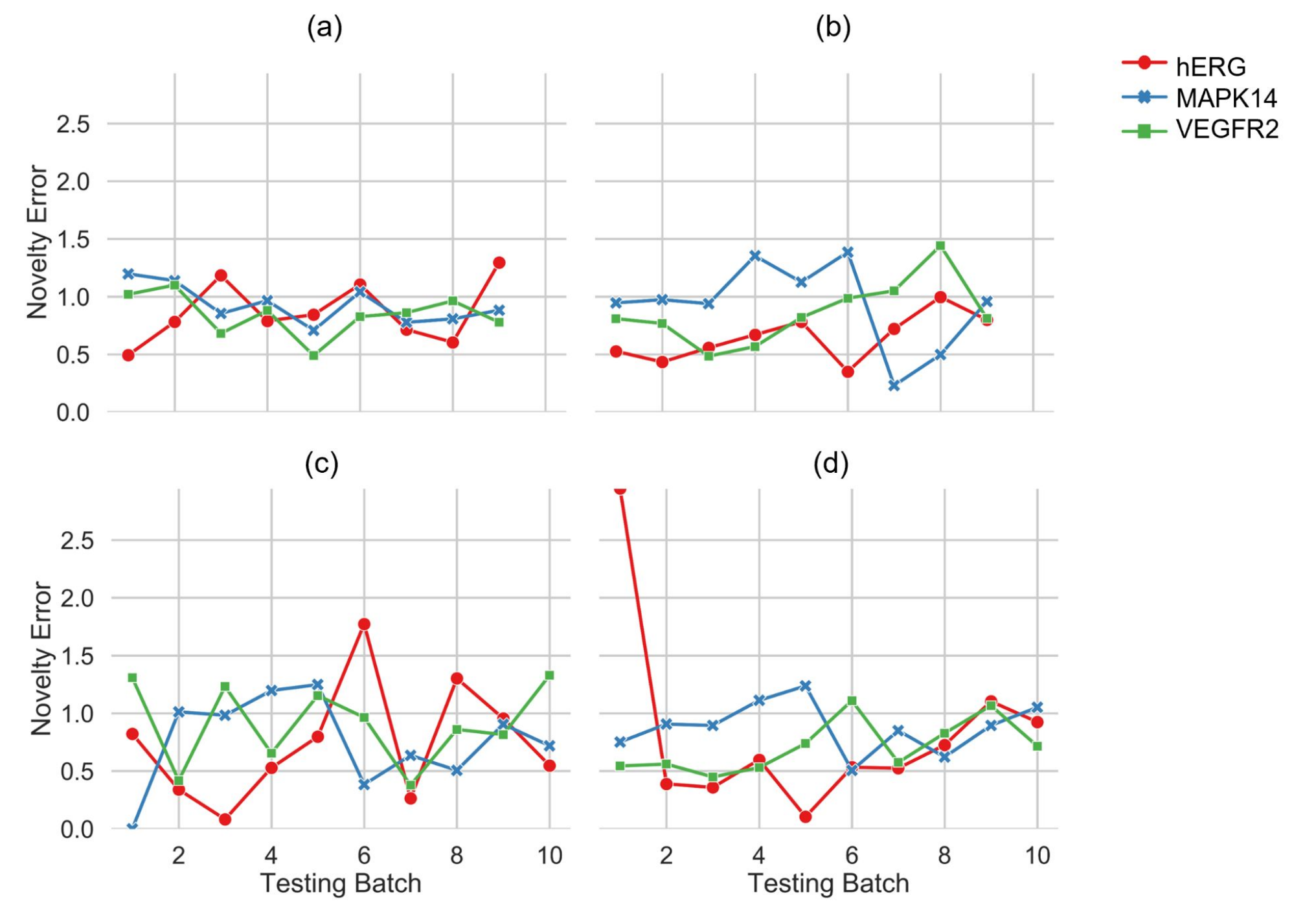

Figure 8: Novelty error for models validated using (a) sorted SFCV, (b) unsorted SFCV, (c) cross-validation with random splits, and (d) cross-validation with scaffold splits for all three protein targets when using a Random Forest model. SFCV: Step-Forward Cross-Validation.

Conclusions: Sorted SFCV provides a systematic approach to validating predictive models in drug discovery, simulating real-world optimization processes. It enhances the ability to predict structurally novel compounds with desirable bioactivity. We recommend incorporating sorted SFCV and evaluating discovery yield and novelty error to better align model testing with the needs of drug discovery pipelines.